White Paper Report

Report ID: 104375

Application Number: HT-50044-11

Project Director: Gregory Crane (gregory.crane@tufts.edu)

Institution: Tufts University

Reporting Period: 10/1/2011-6/30/2015

Report Due: 9/30/2015

Date Submitted: 9/30/2015

Final White Paper

HT-50044-11

"Working with Text in a Digital Age"

Institutes for Advanced Topics in the Digital Humanities (IATDH)

Prof. Gregory R. Crane, Project Director

Tufts University

Submitted 9/30/15

a. Project Activities

During the course of this grant, a three week institute, a two day follow-on workshop and a series of digital humanities lectures were held.

The majority of initial grant activity involved the planning and holding of the NEH Institute in the summer of 2012 entitled Working With Text in a Digital Age. Planning in earnest for this institute began in the fall of 2011 with a general announcement posted on both the Perseus Digital Library[1] website and distributed via relevant topical listservs. This initial publicity was then followed up with the creation of a website[2] with a detailed request for proposals (RFP) in January 2012. Ultimately, the institute received over 80 applications as well as 15 inquiries before the deadline. Institute organizers, lecturers, and Perseus Digital Library staff endeavored to select applicants who could form a coherent topical concentration of study for the institute. Organizers also aimed for a mix of domestic and international participants and both established and promising new researchers. After narrowing the first round of applications by these criteria, final selection of participants was based on research agendas as presented in their applications as well as a willingness to publish all data and final materials as open access

In the spring of 2012, 27 initial participants were selected and communication began between participants and organizers concerning the institute curriculum. The spring of 2012 also involved a great deal of planning for the participants arrival on campus including working with the Tufts Conference Bureau to finalize accommodations, meal plans, and other essential requirements and communicating with participants regarding travel details, reimbursement, visa details (for international participants) and technical requirements. As the logistics for participants were being finalized, the project directors created the final syllabus and course materials. Participants were kept informed of all updates and information at the project website.

The NEH Institute: "Working With Text in a Digital Age" was held from July 23, 2012 to August 24, 2012, with 24 participants attending full time. In addition to the in-residence attendees, a number of local participants including Tufts professors and librarians from Tisch Library were also able to join the workshop for specific lectures or sessions.

[1] http://www.perseus.tufts.edu

[2] http://sites.tufts.edu/digitalagetext/

The first week of the institute involved presentations by the project directors and senior software development staff at Perseus. Topics included corpus linguistics and design, the creation of customized digital editions,[3] and text reuse and detection. During the second week, guest lecturers taught an extensive course on statistics and the use of R that allowed for a great deal of hands on work by participants. Due to the high level of participant interest, there were also two guest specialist Skype lectures on the topic of text reuse and on adapting Optical Character Recognition (OCR) to historical languages. The third week was split between participants presenting on their research interests and work conducted at the institute in the morning and guest lectures on various special topics in the afternoon including: student involvement in digital editing of primary sources, topic modeling, and historical language instruction and learning in a digital world. The last day of the institute (half day) was dedicated to a larger discussion of the digital humanities and digital editing and gave participants an opportunity to provide feedback on the institute as well. Several local area professors and librarians participated as walk-on guests in this session.

A course website was created for the institute using Tufts open source courseware Trunk.[4] The password-protected website was used for communication with participants throughout the institute as well to host the final syllabus, provide access to materials and suggested readings, and compile links to resources. The Trunk site was also used by participants throughout the institute to communicate with each other, organize social outings, and to communicate various research and technical problems. All final presentations by participants were also required to be loaded onto the Trunk site as well as related materials for download.

With the conclusion of the institute, the next major goal was to promote the results of the institute and to plan a follow-on workshop. Some initial outreach and promotion of the planned follow up workshop was conducted at the inaugural conference of the Digital Classics Association: “Word, Space, Time: Digital Perspectives on the Classical World”[5] held at the University of Buffalo on April 5-6, 2013. Despite this initial outreach and planning for the workshop, these preliminary plans needed to be altered as a significant scholarly and research opportunity became available to the project director in the spring of 2013 when he was named to a Humboldt Professorship at the University of Leipzig, Germany. While the original work plan had specified that the two day workshop would be held in the summer of 2013, these changes necessitated the moving of the follow-on workshop, which was thus held in spring 2014.

Consequently, an announcement for the 2014 workshop[6] entitled “Publishing Text for a Digital Age” was made on the institute website. This workshop focused on “establishing specific guidelines for digital publications that publish and or annotate textual sources from the human record” and encouraged submissions of either collections of data with narrative subscriptions or contributions and new digital forms of publication. The institute directors also encouraged 2012 participants to serve on the program committee.

[3] https://github.com/TuftsUniversity/tei-digital-age
[4] https://trunk.tufts.edu/portal/site/digitalagetext [Note: organizers may create a guest login upon request.]
[5] http://classics.buffalo.edu/events/dcaconference/
[6] http://sites.tufts.edu/digitalagetext/2014-workshop/

The workshop “Publishing Text for a Digital Age”[7] was held on March 28-29, 2014, with a diverse array of presentations on topics including the development of digital grammars and other reference tools for classical languages, the creation of digital editions, and how to successfully publish this data as open access. Presentations were held across both days and the workshop ended with a plenary discussion of the different types of digital publication available for historical language materials and their various scholarly accompaniments (apparatus criticus, commentaries, annotations, etc.). Although smaller in scale than originally anticipated, this workshop allowed the institute to further discussion regarding many challenges addressed by the 2012 institute including the difficulties of adapting open source OCR and other document recognition tools for historical languages, how to support best practices in open access publication for classical language materials, and how to best involve students in the work of digital humanities scholarship.

With the remaining funds on the award, the project was able to host a series of digital humanities lectures with a proposed focus on the types of methods and techniques that were taught during the NEH Institute and historical language materials. Thus a call for papers for “Digital Classicist New England”[8](DCNE) was issued on September 1, 2014. Although entitled “Digital Classicist” this was largely due to the inspiration for the series by London’s Digital Classicist Work in Progress Seminar[9] and the call for papers noted that contributions were sought for research being conducted beyond the Greco-Roman world.

After submissions were collected and reviewed in late 2014, the series was announced and promoted at the beginning of 2015. The 10 seminars[10] in total ran from January to April 2015 at Tufts, College of the Holy Cross, Brandeis and Northeastern University. Each seminar was recorded and access to the abstracts, slides, and videos can all be found at the DCNE website. Some lectures were presented by the scholars in person while a number of international scholars made presentations remotely. Topics ranged from presenting digital texts and commentaries on authors such as Vergil and inscriptions online in new ways, how to build open access scholarly communities, social network analysis for classical texts, the development of new visualization methods for archaeological sites, and cutting edge OCR tools for historical languages

b. Accomplishments

This project largely accomplished its most important purpose as stated in the original proposal’s statement of significance, in that it sought to challenge “humanists to spend three weeks not only to acquire new skills but also to transform the way in which they conceive of their research and teaching.” Institute participants were not only introduced to a large number of different technologies in their three weeks but were also challenged to apply those methods to their own

[7] A full schedule can be found here: http://sites.tufts.edu/digitalagetext/2014-workshop/schedule/
[8] http://sites.tufts.edu/perseusupdates/?p=412
[9] http://www.digitalclassicist.org/wip/
[10] A full list of the presentations and related materials can be found here https://sites.tufts.edu/perseusupdates/events/dcne.

work. The final attendee presentations and the continuing number of institute and workshop participants who continue to work in the digital classics and digital humanities fields. indicate that the institute met this goal.

As outlined in the proposal narrative, the institute sought to explore four questions in particular: how technologies would affect participants research agendas, how everyone's research could advance and draw upon intellectual life beyond the academy such as through crowdsourcing or wikis, how technologies would affect formal teaching conducted by participants and in particular open up new opportunities for undergraduate education, and how easily could participants begin to work with more languages so that their ideas could operate within a global network of scholars.

Over the course of the institute there was a great deal of time spent on all of these questions, albeit with less of a focus on how to make use of crowdsourcing and citizen scholars than perhaps originally intended. Nonetheless, the participants all explored new techniques over the course of the institute, presented on how the use of these methods would — or in some cases would not — change their research, and strategized as to how to bring these new methods into the classroom. Presentations at the institute that demonstrated collaborative digital editing tools such as Perseids[11] and online language learning tools were particularly popular. In addition, as the institute included scholars and students from South America, China, Europe as well as the United States and worked with a range of historical languages (Greek, Latin, Old English, Persian, Arabic, Coptic, etc.) participants indeed explored not only sharing their ideas across different modern languages but also discovered how many digital tools and methods for a different historical language might be adapted to the language with which they worked.

While the originally proposed curriculum and format changed somewhat in that the first week of the institute did not dedicate one day exclusively to a particular technology or set of technologies (e.g. Day 1 of Week 1 proposed to cover only structural markup and linguistic annotation, Day 2, Data Visualization, etc.), the basic topic of the workshop remained unchanged. A more detailed review of the research interests and background of the participants led to the project directors deciding on a different approach to the course. A broad exposure to many technologies was deemed to be less desirable than a more *focused* introduction to techniques more applicable to participant research interests, concentrated as they were around some common themes.

Consequently, the first week heavily focused on corpus linguistics, corpus design, annotation types and tools, and the design of digital editions using TEI-XML,[12] with a smaller focus on specific research questions such as text-reuse and intertextuality. While originally the second week had been planned to be exclusively dedicated to project design and development, the experience of the first week made clear that participants had a great desire to learn more about statistics and visualization. As a result, the instruction during week 2 focused heavily on

[11] http://perseids.org/
[12] http://www.tei-c.org/

teaching the use of R[13] as a specific tool with time set aside for individuals and groups to work together on their research questions. Originally, the organizers had planned on assigning participants to working groups to support collaborative work but these working groups organically formed around research questions and topics. The third week followed the format outlined in the proposal with participant presentations and special guest speakers topically organized by day so that the speakers and participants could make the most of the discussions.

Perhaps the most significant objective outlined in the grant and not fully accomplished by the project was the desire to publish not only the data created by institute and workshop participants but also any final presentations and publications based upon that data. Initially the institute organizers had proposed to publish and preserve any final data and publications in either the Tufts digital repository or the Perseus Digital Library. As the types of data created by participants included treebanks, TEI-XML editions, raw OCR files, and corpora utilizing various standards, there was no optimal, comprehensive solution able to encompass the diversity of materials. The aging infrastructure of the current Perseus Digital Library (P4) is not well-positioned to offer the kind of publication platform necessary. The Tufts Digital Library[14] is also currently working towards addressing such diverse data types, but it simply does not have the resources in place to support a broad base of contributors from outside of the Tufts community. The challenges concerning long-term digital preservation and curation of digital humanities data and scholarship— procedural changes, comprehensive data management policies, and the creation of necessary infrastructure — are not unique to Tufts and form the basis of a larger conversation happening at many academic libraries. Nonetheless, the use cases and challenges posed by the institute have become the centerpiece of discussions between the Perseus Digital Library, the Perseids Project, the Tufts Digital Library, and Tufts Research Technology.

c. Audiences

The intended audience for this project was international in scope and had no specific focus on age or gender. The project was directed at specialists in that it required some minimal experience of and interest in the digital humanities. Nonetheless, the project solicited participation not just from faculty but from students at both the graduate and undergraduate level, from library professionals, as well as from those who worked at cultural heritage institutions. The only essential requirements were 1) a clearly defined set of humanities data and 2) a willingness to experiment with digital technologies in order to make that data available as open access. Applicants for the 2012 institute included undergraduate and graduate students, postdocs, junior and senior faculty, library professionals, and independent scholars all of whose work spanned the humanities disciplines.

With over 80 applicants for the initial institute from such a wide range of backgrounds and a final group of participants that included librarians, PhD students, independent researchers and

[13] According to the project website (https://www.r-project.org/), R is a “free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS”

[14] http://dl.tufts.edu/

faculty at all levels, the institute managed to successfully reach a diverse group of individuals interested in the digital humanities. Similarly, both the 2014 workshop and, in particular, the 2015 DCNE lecture series helped the project team to continue this outreach in terms of helping promote the work of both local and international scholars/students who were conducting digital humanities work with historical language materials. The DCNE series of lectures were recorded and all presentations are available on Youtube[15] making them accessible beyond the conclusion of this project.

In sum, over 75 individuals (including project staff, director and co-directors) participated in these activities, including faculty, researchers and students from across the United States with international participants from Belgium, Brazil, Canada, France, Hong Kong, Germany, Israel, Italy, Sweden and the United Kingdom. Faculty and graduate students from both public and private institutions attended or presented as well as a number of independent researchers. A number of the graduate students in attendance have now completed their PhDs and are continuing on in their digital humanities work.

d. Evaluation

There was no formal evaluation conducted of the various activities carried out during the grant period. Initial plans for several rounds of participant surveys were not completed, although the Digital Classics Association provided an informal opportunity for discussions with many institute researchers. In addition, part of the original design of the work plan had specified that the second workshop (held in 2014) could be used, in part, to evaluate the institute's success by benchmarking the progress of institute attendees who participated in the second, follow-on workshop. Although some of the institute attendees submitted or evaluated proposals for the 2014 workshop, overall participation was not large enough to facilitate the intended evaluation.

Over the course of this grant, there were a number of different challenges, both anticipated and unanticipated. The initial response to the 2012 institute was overwhelming and a manageable system to evaluate the various applications had to be developed largely on the fly. This led to the use of the EasyChair conference system for both the 2014 workshop and the 2015 DCNE Lectures. The response for both of these events was modest in comparison to the institute, so this system while useful, was not entirely necessary. This was partially due perhaps to the nature of the events, as the announcement of the 3 week institute generated far more publicity outside of our own announcement on the website than either the workshop or the DCNE series. It may also speak to a desire for extended participation and collaboration as opposed to the more traditional outlet of a short workshop with roundtable discussions or a formal lecture series.

While designing the curriculum and material to be covered for a three week institute was a difficult task in and of itself, adapting that curriculum to meet the needs of participants with very different levels of digital experience also took far more time than originally expected. Although not required for participation, it turned out that participants defined the idea of basic XML or TEI

[15] https://www.youtube.com/channel/UCdif11_Fia_bbzDAAWVBmvA

fluency quite differently and as a consequence one entire day of the first week of the institute was shifted toward teaching a basic boot camp on designing digital editions.

In addition, some of the more challenging if not perhaps new lessons learned from this project included: 1) the difficulty of preserving institute learning materials (originally presented on an open source but still closed access course management system) and making them available and or convenient for access or download in the long-term; 2) the difficulty of continuing momentum after such an event and providing meaningful communication after it has concluded; and 3) the difficulty of publishing and curating new types of born-digital publication in the digital humanities.

To begin with, a number of institute participants expressed the desire for the institute course site on Trunk to be maintained for the indefinite future in order to download materials or alternately asked the institute hosts to perhaps provide another convenient way for participants to download all institute materials. After communication with IT staff it was determined that the Trunk site would remain accessible indefinitely, but at the same time this meant the participants had to log in to the site and download any items one at a time using a rather cumbersome interface. As of this writing, materials created for and by institute participants are still available on the Trunk site but at no other location. In terms of promoting long-term collaboration, plans for the maintenance of a listserv to help institute participants maintain communication between each other were unfortunately never fully carried through, although an initial listserv was setup.

Finally, the original grant proposal included rather ambitious plans to publish the data created by institute participants as well as their final publications — most of which were, in fact, non-traditional in form. One reason that such publication has not yet occurred is that the types of data that participants had at the end of the institute was in varying stages of completeness. Moreover, managing and curating data in the humanities is an ongoing research task faced by both the digital humanities and academic library communities. Most of the participants final presentations documented what they had learned and what they were now going to do differently or experiment with in terms of their data, and this experience partially inspired the project team to alter the format for the 2014 workshop and the DCNE series to encourage people to publish on work already underway, including long-term initiatives.

On final assessment, perhaps one of the greatest successes of this project lies in the relationships and collaborative partnerships formed not only between participants at the institute but also between participants and the hosts of the institute. A number of institute participants became Perseus or Leipzig project staff, went on to work with the related Perseids Project[16], or collaborated independently in their own related research interests. Another key strength of the institute program was that it offered in-depth exploration of a discrete number of challenging methods and techniques from computer science and computational linguistics with examples tailored specifically to participants research goals. In particular, this helped to demonstrate the broad applicability of many of these techniques to different types of corpora as well as materials

[16] http://perseids.org

in different historical languages and this often emboldened participants to try new techniques as well as to see opportunities for partnership and collaboration across disciplines where previously they did not.

e. Continuation of the Project

The work and scholarship conducted during these events and the partnerships that formed continue on in a number of ways. This Institute contributed materially to the award that project director Gregory Crane received to create the Humboldt Chair of Digital Humanities at the University of Leipzig, Germany. Those participants that were interested in the treebanking of historical texts have continued to collaborate both with two related research projects, the Perseids Project and Open Philology, one initiative of the Humboldt Professorship of Digital Humanities at the University of Leipzig.[17]

In addition, original project staff have also continued to explore some of the issues raised during the course of this grant. Institute lecturers and staff are continuing research through the Perseids project into how best to support and preserve digital publications (particularly in the form of student annotations on ancient texts and manuscripts) as well as into how to successfully integrate digital humanities methods into the undergraduate curriculum. Additionally, the Perseids project is heavily involved in the Research Data Alliance (RDA)[18] working group on humanities data that is exploring many of the issues regarding preserving and publishing digital humanities data considered by the institute.

Finally, one new collaborative partnership that was developed during the course of this grant was that between the Perseus Digital Library research project and Tisch Library at Tufts University. Several library staff members attended different parts of the institute in order to explore how they might begin to better support digital humanities scholarship and Perseus project staff have consulted on a variety of library projects including the development of a new Hydra digital repository system to support student and faculty contributions and on the development of use cases and data models to support preservation of digital humanities data.

f. Long Term Impact

The original proposal suggested that its long term impact would involve four concrete results: 1) greater knowledge of TEI-XML for the participants; 2) humanists with an ability to integrate methods from computational linguistics into their textual analysis work; 3) new sources of data produced by participants that could support corpus linguistics; and 4) more humanists who understand how to integrate teaching and research.

Many of the original institute participants have gone on to complete their PhDs and completed their dissertations as digital editions. In addition, some institute participants have gone on to incorporate digital humanities courses as well as specific techniques such as treebanking or the

[17] http://www.dh.uni-leipzig.de/wo/open-philology-project/

[18] https://rd-alliance.org/groups/digital-practices-history-and-ethnography-ig.html. This working group seeks to “advance data standards, practices and infrastructure for historical and ethnographic research, contributing to broader efforts in the digital humanities and social sciences.”

us of R as part of their regular teaching. Similarly, the instructional coursework at the Digital Humanities program at Leipzig has been informed by the themes explored at the various events during the course of this grant, such as digital scholarly editing, humanities programming, citizen science in the humanities, digital philology, and annotation of linguistic sources.[19]

Finally, one notable long term collaboration formed at the institute is that of Caroline Schroeder and Amir Zeldes, who have received three grants from the NEH, a 2014 Digital Humanities start-up grant[20] to develop a Coptic corpus and tools to research and annotate it, a Humanities Collections and Reference Resources grant[21] to work on this same Coptic Scriptorium and in 2015 they received a NEH-DFG grant to work on KELLIA, for "Koptische/Coptic Electronic Language and Literature International Alliance," a collaboration among Pacific University, Georgetown University and two universities in Germany.

g. Grant Products

The only products that resulted during the course of this project were a series of website designed to provide instructions, schedules of events, and provide links to participant abstracts and presentations.

http://sites.tufts.edu/digitalagetext/ — The project website for 2012 insitute and 2014 workshop.
https://github.com/TuftsUniversity/tei-digital-age — The website for the one day tutorial on creating digital editions.
https://trunk.tufts.edu/portal/site/digitalagetext — Course website for the 2012 Institute.
https://sites.tufts.edu/perseusupdates/events/dcne — Website for the DCNE Lecture series with links to abstracts and presentations.
https://www.youtube.com/channel/UCdif11_Fia_bbzDAAWVBmvA — Youtube channel for DCNE.

[19] http://www.dh.uni-leipzig.de/wo/courses/
[20] https://securegrants.neh.gov/PublicQuery/main.aspx?f=1&gn=HD-51907-14
[21] https://securegrants.neh.gov/PublicQuery/main.aspx?f=1&gn=PW-51672-14